OPEN ACCESS

Original research

Genetic architectures of proximal and distal colorectal cancer are partly distinct

Jeroen R Huyghe ,[1] Tabitha A Harrison,[1] Stephanie A Bien,[1] Heather Hampel,[2] Jane C Figueiredo,[3,4] Stephanie L Schmit,[5] David V Conti,[6] Sai Chen,[7] Conghui Qu,[1] Yi Lin,[1] Richard Barfield,[1] John A Baron,[8] Amanda J Cross,[9] Brenda Diergaarde,[10,11] David Duggan,[12] Sophia Harlid,[13] Liher Imaz,[14] Hyun Min Kang,[7] David M Levine,[15] Vittorio Perduca,[16,17] Aurora Perez-Cornago,[18] Lori C Sakoda,[1,19] Fredrick R Schumacher,[20] Martha L Slattery,[21] Amanda E Toland,[22] Fränzel J B van Duijnhoven,[23] Bethany Van Guelpen,[13] Antonio Agudo,[24] Demetrius Albanes,[25] M Henar Alonso,[26,27,28] Kristin Anderson,[29] Coral Arnau-Collell,[30] Volker Arndt,[31] Barbara L Banbury,[1] Michael C Bassik,[32] Sonja I Berndt,[25] Stéphane Bézieau,[33] D Timothy Bishop,[34] Juergen Boehm,[35] Heiner Boeing,[36] Marie-Christine Boutron-Ruault,[17,37] Hermann Brenner ,[31,38,39] Stefanie Brezina,[40] Stephan Buch,[41] Daniel D Buchanan ,[42,43,44] Andrea Burnett-Hartman,[45] Bette J Caan,[46] Peter T Campbell,[47] Prudence R Carr,[48] Antoni Castells,[30] Sergi Castellví-Bel,[30] Andrew T Chan ,[49,50,51,52,53,54] Jenny Chang-Claude,[55,56] Stephen J Chanock,[25] Keith R Curtis,[1] Albert de la Chapelle,[57] Douglas F Easton,[58] Dallas R English,[42,59] Edith J M Feskens,[23] Manish Gala,[49,51] Steven J Gallinger,[60] W James Gauderman,[6] Graham G Giles,[42,59] Phyllis J Goodman,[61] William M Grady,[62,63] John S Grove,[64] Andrea Gsur ,[40] Marc J Gunter,[65] Robert W Haile,[4] Jochen Hampe ,[41] Michael Hoffmeister ,[31] John L Hopper,[42,66] Wan-Ling Hsu,[15] Wen-Yi Huang ,[25] Thomas J Hudson,[67] Mazda Jenab ,[65] Mark A Jenkins,[42] Amit D Joshi,[51,53] Temitope O Keku,[68] Charles Kooperberg,[1] Tilman Kühn,[55] Sébastien Küry,[33] Loic Le Marchand,[64] Flavio Lejbkowicz,[69,70,71] Christopher I Li,[1] Li Li,[72] Wolfgang Lieb,[73] Annika Lindblom,[74,75] Noralane M Lindor,[76] Satu Männistö,[77] Sanford D Markowitz,[78] Roger L Milne,[42,59] Lorena Moreno,[30] Neil Murphy ,[65] Rami Nassir,[79] Kenneth Offit,[80,81] Shuji Ogino,[52,53,82,83] Salvatore Panico,[84] Patrick S Parfrey,[85] Rachel Pearlman,[2] Paul D P Pharoah,[58] Amanda I Phipps,[1,86] Elizabeth A Platz,[87] John D Potter,[1] Ross L Prentice,[1] Lihong Qi,[88] Leon Raskin,[89] Gad Rennert,[70,71,90] Hedy S Rennert,[70,71,90] Elio Riboli,[91] Clemens Schafmayer,[92] Robert E Schoen,[93] Daniela Seminara,[94] Mingyang Song,[49,51,95] Yu-Ru Su,[1] Catherine M Tangen,[61] Stephen N Thibodeau,[96] Duncan C Thomas,[6] Antonia Trichopoulou,[97,98] Cornelia M Ulrich,[35] Kala Visvanathan,[87] Pavel Vodicka,[99,100,101] Ludmila Vodickova,[99,100,101] Veronika Vymetalkova,[99,100,101] Korbinian Weigl,[31,39,102] Stephanie J Weinstein,[25] Emily White,[1] Alicja Wolk,[103] Michael O Woods,[104] Anna H Wu,[6] Goncalo R Abecasis,[7] Deborah A Nickerson,[105] Peter C Scacheri,[106] Anshul Kundaje,[32,107] Graham Casey,[108] Stephen B Gruber,[109,110] Li Hsu,[1,15] Victor Moreno,[26,27,28] Richard B Hayes,[111] Polly A Newcomb,[1,86] Ulrike Peters [1,86]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/gutjnl-2020-321534).

For numbered affiliations see end of article.

Correspondence to
Dr Ulrike Peters, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; upeters@fhcrc.org

Albert de la Chapelle is deceased.

Received 23 April 2020
Revised 26 November 2020
Accepted 18 December 2020

Check for updates

To cite: Huyghe JR, Harrison TA, Bien SA, *et al*. *Gut* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/gutjnl-2020-321534

ABSTRACT

Objective An understanding of the etiologic heterogeneity of colorectal cancer (CRC) is critical for improving precision prevention, including individualized screening recommendations and the discovery of novel drug targets and repurposable drug candidates for

Significance of this study

What is already known on this subject?

- Heterogeneity among colorectal cancer (CRC) tumours originating at different locations of the colorectum has been revealed in somatic genomes, epigenomes and transcriptomes, and in some established environmental risk factors for CRC.
- Genome-wide association studies (GWASs) have identified over 100 genetic variants for overall CRC risk; however, a comprehensive analysis of the extent to which genetic risk factors differ by the anatomical sublocation of the primary tumour is lacking.

What are the new findings?

- In this large consortium-based study, we analysed clinical and genome-wide genotype data of 112 373 CRC cases and controls of European ancestry to comprehensively examine whether CRC case subgroups defined by anatomical sublocation have distinct germline genetic aetiologies.
- We discovered 13 new loci at genome-wide significance ($p<5\times10^{-8}$) that were specific to certain anatomical sublocations and that were not reported by previous GWASs for overall CRC risk; multiple lines of evidence support strong candidate target genes at several of these loci, including *PTGER3*, *LCT*, *MLH1*, *CDX1*, *KLF14*, *PYGL*, *BCL11B* and *BMP7*.
- Systematic heterogeneity analysis of genetic risk variants for CRC identified thus far, revealed that genetic architectures of proximal and distal CRC are partly distinct, and demonstrated that distal colon and rectal cancer have very similar germline genetic aetiologies.
- Taken together, our results further support the idea that tumours arising in different anatomical sublocations of the colorectum may have distinct aetiologies.

How might it impact on clinical practice in the foreseeable future?

- Our results provide an informative resource for understanding the differential role that genetic variants, genes and pathways may play in the mechanisms of proximal and distal CRC carcinogenesis.
- The new insights into the aetiologies of proximal and distal CRC may inform the development of new precision prevention strategies, including individualised screening recommendations and the discovery of novel drug targets and repurposable drug candidates for chemoprevention.
- Our findings suggest that future studies of aetiological risk factors for CRC and molecular mechanisms of carcinogenesis should take into consideration the anatomical sublocation of the colorectal tumour. In particular, our results argue against lumping proximal and distal colon cancer cases.

chemoprevention. Known differences in molecular characteristics and environmental risk factors among tumors arising in different locations of the colorectum suggest partly distinct mechanisms of carcinogenesis. The extent to which the contribution of inherited genetic risk factors for CRC differs by anatomical subsite of the primary tumor has not been examined.

Design To identify new anatomical subsite-specific risk loci, we performed genome-wide association study (GWAS) meta-analyses including data of 48 214 CRC cases and 64 159 controls of European ancestry. We characterised effect heterogeneity at CRC risk loci using multinomial modelling.

Results We identified 13 loci that reached genome-wide significance ($p<5\times10^{-8}$) and that were not reported by previous GWASs for overall CRC risk. Multiple lines of evidence support candidate genes at several of these loci. We detected substantial heterogeneity between anatomical subsites. Just over half (61) of 109 known and new risk variants showed no evidence for heterogeneity. In contrast, 22 variants showed association with distal CRC (including rectal cancer), but no evidence for association or an attenuated association with proximal CRC. For two loci, there was strong evidence for effects confined to proximal colon cancer.

Conclusion Genetic architectures of proximal and distal CRC are partly distinct. Studies of risk factors and mechanisms of carcinogenesis, and precision prevention strategies should take into consideration the anatomical subsite of the tumour.

INTRODUCTION

Despite improvements in prevention, screening and therapy, colorectal cancer (CRC) remains one of the leading causes of cancer-related death worldwide, with an estimated 53 200 fatal cases in 2020 in the USA alone.[1] CRCs that arise proximal (right) or distal (left) to the splenic flexure differ in age-specific and sex-specific incidence rates, clinical, pathological and tumour molecular features.[2–5] These observed differences reflect a complex interplay between differential exposure of colorectal crypt cells to local environmental carcinogenic and protective factors in the luminal content (including the microbiome), and distinct inherent biological characteristics that may influence neoplasia risk, including sex and differences between anatomical segments in embryonic origin, development, physiology, function and mucosal immunology. The precise extrinsic and intrinsic aetiological factors involved, their relative contributions, and how they interact to influence the carcinogenic process remain largely elusive.

An individual's genetic background plays an important role in the initiation and development of CRC. Based on twin registries, heritability is estimated to be around 35%.[6] Since genome-wide association studies (GWASs) became possible just over a decade ago, over 100 independent common genetic variant associations for overall CRC risk have been identified, over half of which were identified in the past few years.[7–10] Three decades ago, based on observed similarities between Lynch syndrome and proximal CRC, and between familial adenomatous polyposis and distal CRC, Bufill proposed the existence of two distinct genetic categories of CRC according to the location of the primary tumour.[2] However, given that genetic variants that influence CRC risk typically have small effect sizes, until very recently, sample sizes did not provide adequate statistical power to conduct meaningful subsite analyses. As a consequence, GWASs to detect genetic associations specific to CRC case subgroups defined by primary tumour anatomic subsite have not been reported yet. Similarly, a comprehensive analysis of the extent to which allelic risk of known GWAS-identified variants differs by primary tumour anatomic subsite is lacking.

To address the major gap in our knowledge of the differential role that genetic variants, genes and pathways play in mechanisms of proximal and distal CRC carcinogenesis, we analysed clinical and genome-wide genotype data for 112 373 CRC cases and controls. First, to discover new loci and genetic risk variants with site-specific allelic effects, we conducted GWASs of case subgroups defined by the location of their primary tumour within the colorectum. Next, we systematically characterised heterogeneity of allelic effects between primary tumour subsites for new and previously identified CRC risk variants to identify loci with shared and site-specific allelic effects.

METHODS

Detailed methods are provided in online supplemental materials.

Samples and genotypes

This study included clinical and genotype data for 48 214 CRC cases and 64 159 controls from three consortia: Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), Colorectal Cancer Transdisciplinary Study (CORECT) and Colorectal Cancer Family Registry (CCFR). Online supplemental table 1 provides details on sample numbers and demographic characteristics by study. All study participants were of genetically inferred European-ancestry. Across studies, participant recruitment occurred between the early 1990s and the 2010s. Details of genotype data sets, genotype QC, sample selection and studies included in this analysis have been published previously.[7 8 11 12] All participants provided written informed consent, and each study was approved by the relevant research ethics committee or institutional review board.

Colorectal tumour anatomic sublocation definitions

We defined proximal colon cancer as any primary tumour arising in the cecum, ascending colon, hepatic flexure or transverse colon; distal colon cancer as any primary tumour arising in the splenic flexure, descending colon or sigmoid colon; and rectal cancer as any primary tumour arising in the rectum or rectosigmoid junction. For the GWAS discovery analyses, we analysed five case subgroups based on primary tumour sublocation. In addition to the three afore-mentioned mutually exclusive case sets (proximal colon, distal colon and rectal cancer), we defined colon cancer and distal/left-sided colorectal cancer case sets. Colon cancer cases comprised combined proximal colon and distal colon cancer cases, and additional colon cases with unspecified site. In the distal/left-sided colorectal cancer cases analysis, we combined distal colon and rectal cancer cases based on the different embryonic origins of the proximal colon versus the distal colon and rectum. Online supplemental figure 1 and table 1 summarise distributions of age of diagnosis by sex and primary tumour site.

Statistical analysis

GWAS meta-analyses

We imputed all genotype datasets to the Haplotype Reference Consortium panel.[13] In brief, we phased all genotyping array data sets using SHAPEIT2[14] and used the Michigan Imputation Server[15] for imputation. Within each dataset, variants with an imputation accuracy $r^2 \geq 0.3$ and minor allele count ≥ 50 were tested for association with CRC case subgroup. Variants that only passed filters in a single dataset were excluded. We assumed an additive model using imputed genotype dosage in a logistic regression adjusted for age, sex and study or genotyping project-specific covariates, including principal components to adjust for population structure. Details of covariate corrections have been published previously.[8] Because Wald tests can be anticonservative for rare variants, we performed likelihood ratio tests and combined association summary statistics across sample sets via fixed-effects meta-analysis employing Stouffer's method, implemented in the METAL software.[16] Reported p values are based on this analysis. Reported combined OR estimates and 95% CIs are based on an inverse variance-weighted fixed-effects meta-analysis.

Heterogeneity in allelic effect sizes between tumour anatomic sublocations

To characterise tumour subsite-specificity and effect size heterogeneity across tumour subsites for new loci, and for established loci for overall CRC, we examined association evidence in three different ways. First, for each index variant we created forest plots of OR estimates from GWAS meta-analyses for proximal colon, distal colon and rectal cancer. Second, we tested for heterogeneity using multinomial logistic regression. In brief, after pooling of datasets, we performed a likelihood ratio test comparing a model in which ORs for the risk variant were allowed to vary across tumour subsites, to a model in which ORs were constrained to be the same across tumour subsites. Third, inspired by reference,[17] we used a multinomial logistic regression-based model selection approach to assess which configuration of tumour subsites is most likely to be associated with a given variant. For each variant, we defined and fitted 11 possible causal risk models specifying variant effect configurations that vary or are constrained to be equal among subsets of tumour subsites (online supplemental table 2). We then identified and report the best fitting model using the Bayesian information criterion (BIC). For each model i we calculated $\Delta BIC_i = BIC_i - BIC_{min}$, where BIC_{min} is the BIC value for the best model. Models with $\Delta BIC_i \leq 2$ were considered to have substantial support and indistinguishable from the best model.[18] For these variants, we do not report a single best model. Analyses were carried out using the VGAM R package.[19] The list of index variants for previously published CRC risk signals is based on Huyghe *et al*.[8]

Pathway enrichment analyses

We used the Pascal programme to compute pathway enrichment score p values from genome-wide summary statistics.[20] The gene set library used comprises the combined KEGG,[21] REACTOME[22] and BIOCARTA[23] databases.

Genomic annotation of new GWAS loci and gene prioritisation

We annotated all new loci with five types of functional and regulatory genomic annotations: (i) cell-type-specific regulatory annotations for histone modifications and open chromatin, (ii) nonsynonymous coding variation, (iii) evidence of transcription factor binding, (iv) predicted functional impact across different databases, (v) colocalisation with expression quantitative trait loci (eQTL) signals. Genes were further prioritised based on biological relevance, colorectal tissue expression, presence of associated non-synonymous variants predicted to be deleterious, evidence from functional studies, somatic alterations or familial syndromes. Details are in online supplemental materials.

RESULTS

The final analyses included data for 48 214 CRC cases and 64 159 controls of European ancestry. To discover new loci and genetic risk variants with site-specific allelic effects, we conducted five genome-wide association scans of case subgroups defined by the location of their primary tumour within the colorectum: proximal colon cancer (n=15 706), distal colon cancer (n=14 376), rectal cancer (n=16 212), colon cancer, in which we omitted rectal cancer cases (n=32 002), and distal/left-sided CRC, in which we combined distal colon and rectal cancer cases (n=30 588). Next, we systematically characterised heterogeneity of allelic effects between tumour subsites for new and previously identified CRC risk variants to identify loci with shared and site-specific allelic effects.

New colorectal cancer risk loci

Across the five CRC case subgroup GWAS meta-analyses, a total of 11 947 015 single nucleotide variants (SNVs) were analysed. Inspection of genomic control inflation factors and quantile–quantile plots of test statistics indicated no residual population stratification issues (online supplemental materials and figure 2). Across tumour subsites, we identified 13 loci that mapped outside regions previously implicated by GWASs for overall CRC risk (closest known locus 3.1 megabases away) and that reached genome-wide significance ($p<5\times10^{-8}$) in at least one of the meta-analyses (table 1, figure 1, online supplemental figures 3 and 4). Seven of the new loci passed a Bonferroni-adjusted genome-wide significance threshold correcting for five case subgroups analysed (table 1). All lead variants were well imputed (minimum average imputation $r^2=0.788$), had minor allele frequency (MAF) >1%, and displayed no significant heterogeneity between sample sets (Cochran's Q heterogeneity test p>0.05; table 1).

The novel associations showing the strongest statistical evidence were obtained for proximal colon cancer and mapped near *MLH1* on 3p22.2 (rs1800734, $p=3.8\times10^{-18}$) and near *BCL11B* on 14q32.2 (rs80158569, $p=8.6\times10^{-11}$). These loci showed strongly proximal cancer-specific associations. The proximal colon analysis also yielded a locus on 14q32.12 (rs61975764, $p=2.8\times10^{-8}$) that showed attenuated effects for other tumour subsites (figure 1 and online supplemental table 3). Most new loci (six) were discovered in the left-sided CRC analysis: 2q21.3 (rs1446585, $p=3.3\times10^{-8}$), near *CDX1* on 5q32 (rs2302274, $p=4.9\times10^{-9}$), near *KLF14* on 7q32.3 (rs73161913, $p=1.3\times10^{-9}$), 10q23.31 (rs7071258, $p=8.4\times10^{-9}$), 19p13.3 (rs62131228, $p=2.4\times10^{-8}$) and near *BMP7* on 20q13.31 (rs6014965, $p=4.5\times10^{-9}$). The rectal cancer analysis identified an additional locus near *PYGL* on 14q22.1 (rs28611105, $p=4.7\times10^{-9}$) that showed an attenuated effect for distal colon cancer (figure 1 and online supplemental table 3). No additional new loci were detected in the distal colon analysis. The colon cancer analysis identified three new loci: near *PTGER3* on 1p31.1 (rs3124454, $p=1.4\times10^{-8}$), 3p21.2 (rs353548, $p=1.3\times10^{-8}$) and 22q13.31 (rs736037, $p=2.8\times10^{-8}$).

Genomic annotations and most likely target gene(s) at new loci

To gain insight into molecular mechanisms underlying new association signals, and to identify candidate causal variants and target gene(s), we annotated signals with functional and regulatory genomic annotations, assessed colocalisation with eQTLs, and performed literature-based gene prioritisation. Results for all new signals are given in online supplemental tables 4 and 5, and candidate target genes are also given in table 1. Notable and strong candidate target genes include *PTGER3*, *LCT*, *MLH1*, *CDX1*, *KLF14*, *PYGL*, *RIN3*, *BCL11B* and *BMP7*. Strong candidate causal variants were identified at loci 2q21.3 (rs4988235; *LCT*), 3p22.2 (rs1800734; *MLH1*), 14q32.12 (rs61975764; *RIN3*) and 14q32.3 (rs80158569; *BCL11B*). A detailed interpretation of candidate causal variants and target genes is deferred to the Discussion section.

Risk heterogeneity between tumour anatomical sublocations

Multinomial logistic regression modelling of 96 known and 13 newly identified risk variants showed the presence of substantial risk heterogeneity between cancer in the proximal colon, distal colon and rectum. For 61 variants, the heterogeneity p value (p_{het}) was not significant ($p_{het}>0.05$). For 51 of those variants,

Table 1 New genome-wide significant colorectal cancer risk loci identified by genome-wide association analysis of case subgroups defined by primary tumour anatomic subsite

Tumour site*	Locus	Nearby gene(s)	rsID lead variant	Chr.	Position (build 37)	Alleles (risk/ other)	RAF (%)	OR	95% CI	P value	r^2	I^2	P_{het}	N cases	N controls
Colon	1p31.1	*PTGER3*	rs3124454	1	71 040 166	G/T	58.1	1.07	1.04 to 1.09	1.4E-08	0.926	6.1	0.38	32 002	64 159
Left-sided	2q21.3	*LCT*	rs1446585	2	136 407 479	G/A	39.9	1.07	1.04 to 1.10	3.3E-08	1.121	43.7	0.11	30 588	64 159
Proximal colon	3p22.2	*MLH1*	rs1800734†	3	37 034 946	A/G	24.7	1.15	1.11 to 1.19	3.8E-18	1.008	43.8	0.11	15 706	64 159
Colon	3p21.2	*STAB1; TLR9; NISCH*	rs353548	3	52 269 491	G/A	95.3	1.15	1.10 to 1.21	1.3E-08	0.975	0	0.48	32 002	64 159
Left-sided	5q32	*CDX1*	rs2302274†	5	149 546 426	G/A	47.8	1.07	1.04 to 1.09	4.9E-09	1.008	3.8	0.39	30 588	64 159
Left-sided	7q32.3	*KLF14; LINC00513*	rs73161913†	7	130 607 779	G/A	94.3	1.16	1.10 to 1.22	1.3E-09	0.975	0	0.79	30 588	64 159
Left-sided	10q23.31	*PANK1; KIF20B*	rs7071258†	10	91 574 624	A/G	21.6	1.08	1.05 to 1.11	8.4E-09	0.993	0	0.71	30 588	64 159
Rectal	14q22.1	*PYGL; NIN; ABHD12B*	rs28611105†	14	51 359 658	G/T	21.5	1.11	1.07 to 1.15	4.7E-09	0.983	50.5	0.07	16 212	64 159
Proximal colon	14q32.12	*RIN3*	rs61975764	14	93 014 929	G/A	55.3	1.08	1.05 to 1.11	2.8E-08	0.987	0	0.71	15 706	64 159
Proximal colon	14q32.2	*BCL11B*	rs80158569†	14	99 782 937	A/G	7.5	1.18	1.12 to 1.24	8.6E-11	0.899	29.9	0.21	15 706	64 159
Left-sided	19p13.3	*STK11; SBNO2*	rs62131228	19	1 157 642	G/A	98.1	1.28	1.17 to 1.40	2.4E-08	0.788	0	0.95	29 632	63 385
Left-sided	20q13.31	*BMP7*	rs6014965†	20	55 831 203	A/G	55.4	1.07	1.04 to 1.09	4.5E-09	0.995	10.5	0.35	30 588	64 159
Colon	22q13.31	*FAM118A; FBLN1*	rs736037	22	45 724 999	A/G	28.6	1.07	1.04 to 1.09	2.8E-08	1.015	0	0.74	32 002	64 159

Lead variant is the most significant variant at the locus. Reference single nucleotide polymorphism (SNP) cluster ID (rsID) based on NCBI dbSNP Build 152. Alleles are on the + strand. All p values reported in this table are from a sample size-weighted fixed-effects meta-analysis of logistic regression-based likelihood-ratio test results. Reported imputation qualities r^2 are effective sample size (N_{eff})-weighted means across the six data sets, where $N_{eff}=4/(1/N_{case}+1/N_{control})$. The I^2 statistic measures heterogeneity on a scale of 0–100%. P_{het} is the p value from Cochran's Q test for heterogeneity.
*Colon: proximal colon+distal colon+colon, unspecified site; left-sided: distal colon+rectal. Details of tumour site definitions including ICD-9 codes are given in the Methods section and online supplemental materials.
†Variant attained Bonferroni-adjusted genome-wide significance (5E-08/5=1E-08), corrected for the number of CRC case subgroups analysed.
Chr., chromosome; CRC, colorectal cancer; RAF, risk allele frequency.

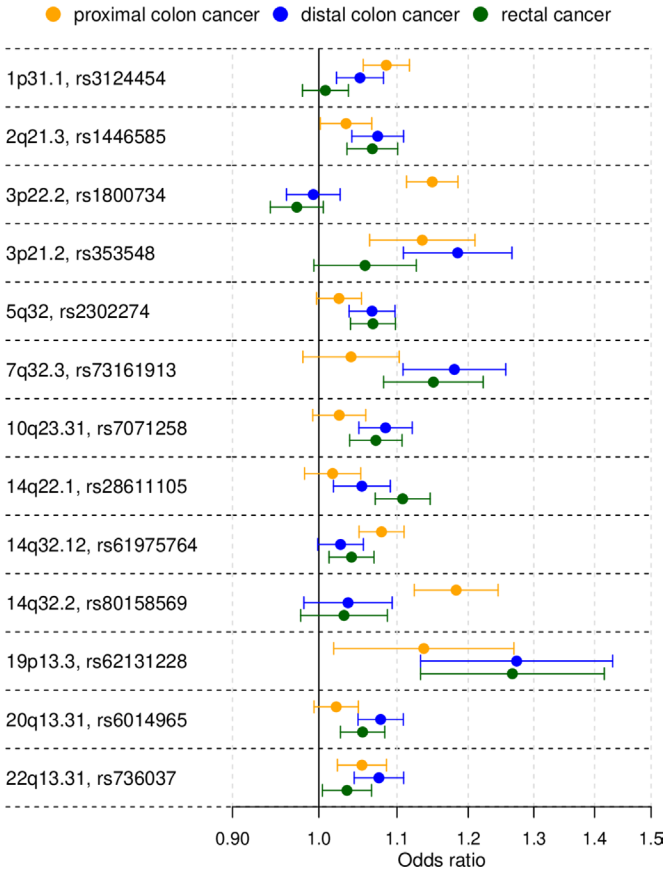

Figure 1 Primary tumour site-specific associations for the lead single nucleotide polymorphisms (SNPs) of the 13 colorectal cancer risk loci not reported in previous genome-wide association studies. The forest plot shows the (log-additive) OR estimates together with 95% CIs. For clarity, this figure only shows results for the proximal colon, distal colon and rectal cancer case subgroup analyses.

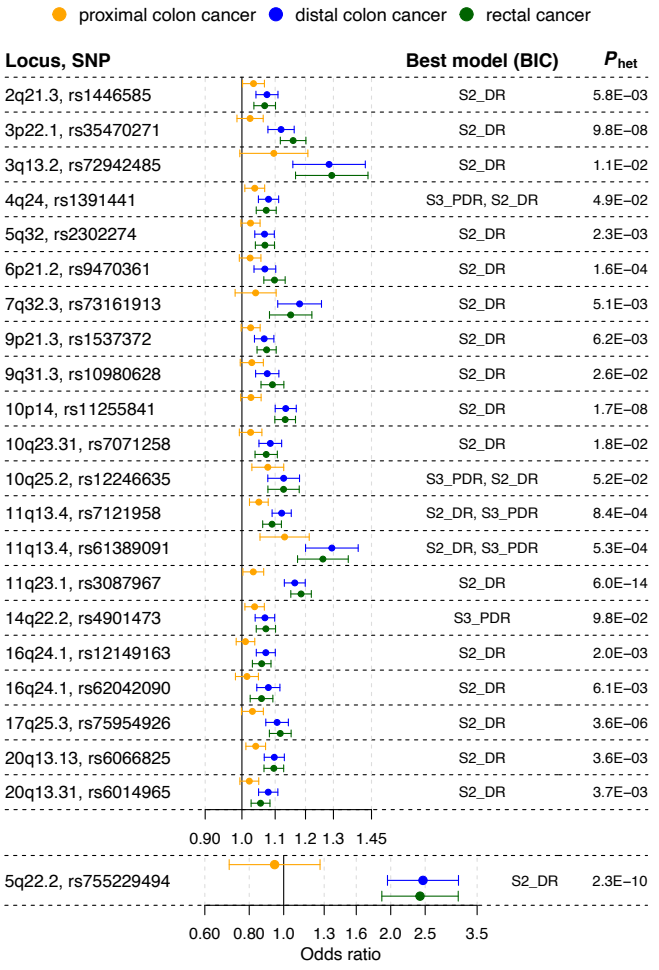

Figure 2 Loci showing association with risk of distal colorectal cancer (ie, distal colon+rectal), but attenuated or no evidence for association with proximal colon cancer risk. The forest plot shows the (log-additive) OR estimates for the lead single nucleotide polymorphisms (SNPs) at the loci, together with 95% CIs, from the genome-wide association study meta-analyses of case subgroups defined by primary tumour anatomical subsite for proximal colon, distal colon and rectal. Best model is the best-fitting multinomial logistic regression model according to the Bayesian information criterion (BIC). Models are defined in online supplemental table 2. P_{het} is the p value from a test for heterogeneity of allelic effects across tumour subsites.

a multinomial model in which ORs were identical for the three cancer sites provided the best fit, and for 8 of the remaining 10 variants, this model did not significantly differ from the best fitting model (online supplemental tables 2, 3 and 7; figure 5).

Among the 109 known or new variants, 48 showed at least some evidence of heterogeneity with $p_{het}<0.05$, and after Holm-Bonferroni correction for multiple testing, 14 variants showing strong evidence of heterogeneity remained significant ($p_{het}<4.6\times10^{-4}$). These included 10 variants previously reported in GWASs for overall CRC risk.

For 17 out of the 48 variants with $p_{het}<0.05$, the best-fitting model supported an effect limited to left-sided CRC (figure 2 and online supplemental tables 3 and 7). Of these 17 variants, 6 were in the list of variants with the strongest evidence of heterogeneity ($p_{het}<4.6\times10^{-4}$), including the following previously reported loci: *C11orf53-COLCA1-COLCA2* on 11q23.1 ($p_{het}=6.0\times10^{-14}$), *APC* on 5q22.2 ($p_{het}=2.3\times10^{-10}$), *GATA3* on 10p14 ($p_{het}=1.7\times10^{-8}$), *CTNNB1* on 3p22.1 ($p_{het}=9.8\times10^{-8}$), *RAB40B-METRLN* on 17q25.3 ($p_{het}=3.6\times10^{-6}$) and *CDKN1A* on 6p21.2 ($p_{het}=1.6\times10^{-4}$). Inspection of forest plots and association evidence also suggest stronger risk effects for left-sided tumours for the following additional five known loci: *TET2* on 4q24, *VTI1A* on 10q25.2, two independent signals near *POLD3* on 11q13.4, and *BMP4* on 14q22.2.

For 5 out of the 49 variants with $p_{het}<0.05$, a model with association with colon cancer risk, but no association with rectal cancer risk, provided the best fit (online supplemental tables 3 and 7). These involve the following loci: *PTGER3* on 1p31.1, *STAB1-TLR9* on 3p21.2, *HLA-B-MICA/B-NFKBIL1-TNF* on 6p21.33, *NOS1* on 12q24.22 and *LINC00673* on 17q24.3. Association evidence also suggests stronger risk effects for colon tumours for one of two independent signals near *PTPN1* on 20q13.13.

Evidence from the three approaches (figure 1; online supplemental tables 3 and 7) indicates that only two loci are strongly proximal colon cancer-specific: *MLH1* on 3p22.2 ($p_{het}=5.4\times10^{-19}$), and *BCL11B* ($p_{het}=1.5\times10^{-5}$) on 14q32.2. Finally, for only one variant, at one of two independent loci near *SATB2* on 2q33.1, a model with a rectal cancer-specific association provided the best fit, but association evidence shows attenuated effects for proximal and distal colon cancer. OR estimates also suggest stronger risk effects for rectal cancer at the known

loci *LAMC1* on 1q25.3, and *CTNNB1* on 3p22.1, and at new locus *PYGL* on 14q22.1.

Pathway enrichment analyses

To explore whether biological pathways play different roles in tumourigenesis of proximal and distal CRC, we conducted pathway enrichment analyses of GWAS summary statistics. There was no clear and strong evidence for differential involvement of pathways; pathways that were Bonferroni-significant for one anatomical subsite, reached at least suggestive significance levels for other subsites (online supplemental table 8). Several of the Bonferroni-significant pathways related to transforming growth factor β (TGFβ) signalling.

DISCUSSION

It has long been recognised that CRCs arising in different anatomical segments of the colorectum differ in age-specific and sex-specific incidence rates, clinical, pathological and tumour molecular features. However, our understanding of the aetiological factors underlying these medically important differences has remained scarce. This study aimed to examine whether the contribution of common germline genetic variants to CRC carcinogenesis differs by anatomical sublocation. The large sample size comprising 112 373 cases and controls provided adequate statistical power to discover new loci and variants with risk effects limited to tumours for certain anatomical subsites, and to compare allelic effect sizes across anatomical subsites.

Our CRC case subgroup meta-analyses identified 13 additional genome-wide significant CRC risk loci that, due to substantial allelic effect heterogeneity between anatomical subsites, were not detected in larger, previously published GWASs for overall CRC risk.[8 9] In fact, the only way to discover certain loci and risk variants with case subgroup-specific allelic effects is via analysis of homogeneous case subgroups.[24] For example, p values for rs1800734 and rs80158569 were ∼18 and∼5 powers of 10, respectively, more significant in the proximal colon analysis compared with in our overall CRC analysis. While follow-up studies are needed to uncover the causal variant(s), biological mechanism and target gene, multiple lines of evidence support strong candidate target genes at many of the new loci, including genes *MLH1*, *BCL11B*, *RIN3*, *CDX1*, *LCT*, *KLF14*, *BMP7*, *PYGL* and *PTGER3*.

At the *MLH1* gene promoter region on 3p22.2, associated to proximal colon cancer, previous studies have reported strong and robust associations between the common single nucleotide polymorphism (SNP) rs1800734, and CRC with high microsatellite instability (MSI-H).[25 26] Rare deleterious nonsynonymous germline mutations in the DNA mismatch repair (MMR) gene *MLH1* are a frequent cause of Lynch syndrome (OMIM #609310). The risk allele of the likely causal SNP rs1800734 is strongly associated with *MLH1* promoter hypermethylation and loss of MLH1 protein in CRC tumours.[26] The mechanisms of *MLH1* promoter hypermethylation and subsequent gene silencing may account for most CRC tumours with defective DNA MMR and MSI-H.[27]

At the highly localised, proximal colon-specific association signal on 14q32.2, lead SNP rs80158569 is located in a colonic crypt enhancer and overlaps with multiple transcription factor binding sites, making it a strong candidate causal variant. Nearby gene *BCL11B* encodes a transcription factor that is required for normal T cell development,[28 29] and that is a SWI/SNF complex subunit.[30] *BCL11B* acts as a haploinsufficient tumour suppressor in T-cell acute lymphoblastic leukaemia.[31 32] Experimental work suggests that impairment of Bcl11b promotes intestinal tumourigenesis in mice and humans through deregulation of the Wnt/β-catenin pathway.[33]

At locus 14q32.12, lead SNP rs61975764 showed the strongest association evidence in the proximal colon analysis and attenuated effects for other tumour locations. Genotype-Tissue Expression (GTEx) data show that rs61975764 is an eQTL for gene Ras and Rab interactor 3 (*RIN3*) in transverse colon tissue. RIN3 functions as a RAB5 and RAB31 guanine nucleotide exchange factor involved in endocytosis.[34 35]

At locus 5q32, associated with left-sided CRC, the intestine-specific transcription factor caudal-type homeobox 1 (*CDX1*) encodes a key regulator of differentiation of enterocytes in the normal intestine and of CRC cells. CDX1 is central to the capacity of colon cells to differentiate and promotes differentiation by repressing the polycomb complex protein BMI1 which promotes stemness and self-renewal. The repression of BMI1 is mediated by microRNA-215 which acts as a target of CDX1 to promote differentiation and inhibit stemness.[36] CDX1 has been shown to inhibit human colon cancer cell proliferation by blocking β-catenin/T-cell factor transcriptional activity.[37]

In a region of extensive LD on locus 2q21.1, lead SNP rs1446585, associated with left-sided CRC, is in strong LD with functional SNP rs4988235 (LD r^2=0.854) in the *cis*-regulatory element of the lactase (*LCT*) gene. In Europeans, the rs4988235 genotype determines the lactase persistence phenotype, or the ability to digest lactose in adulthood. The p value for functional SNP rs4988235 under an additive model was 7.0×10^{-7}. The allele determining lactase persistence (T) is associated with decreased CRC risk. This is consistent with a previously reported association between low lactase activity defined by the CC genotype and CRC risk in the Finnish population.[38] The protective effect conferred by the lactase persistence genotype is likely mediated by dairy products and calcium which are known protective factors for CRC.[39] When we tested for association with left-sided CRC assuming a dominant model, associations for rs1446585 and rs4988235 became more significant with p values of 4.4×10^{-11} and 1.4×10^{-9}, respectively. For functional SNP rs4988235, the OR estimate for having genotype CC versus CT or TT, and left-sided CRC was 1.14 (95% CI 1.09 to 1.19). Because this region has been under strong selection, it is particularly prone to population stratification.[40] However, we adjusted for genotype principal components, and the association showed a consistent direction of effect across sample sets (online supplemental table 6), suggesting this association is not spurious.

Candidate genes at left-sided CRC loci 7q32.2 and 20q13.31 are involved in TGFβ signalling. At 7q32.3, gene Krüppel-like factor 14 (*KLF14*) is a strong candidate. We previously reported loci at known CRC oncogene *KLF5* and at *KLF2*.[8] The imprinted gene *KLF14* shows monoallelic maternal expression, and is induced by TGFβ to transcriptionally corepress the TGFβ receptor 2 (*TGFBR2*) gene.[41] A *cis*-eQTL for *KLF14*, uncorrelated with our lead SNP rs73161913, acts as a master regulator related to multiple metabolic phenotypes,[42 43] and a nearby independent variant is associated to basal cell carcinoma.[44] For both reported associations, effects depended on parent-of-origin of risk alleles. The association with metabolic phenotypes also depended on sex. We did not find evidence for strong sex-dependent effects (men: OR=1.13, 95% CI 1.07 to 1.20; women: OR=1.17, 95% CI 1.09 to 1.25). Further investigation is warranted to analyse parent-of-origin effects. At 20q13.31, gene bone morphogenetic protein 7 (BMP7) is a strong candidate. BMP7 signalling in *TGFBR2*-deficient stromal cells promotes epithelial carcinogenesis through SMAD4-mediated signalling.[45] In CRC tumours,

BMP7 expression correlates with parameters of pathological aggressiveness such as liver metastasis and poor prognosis.[46]

On 14q22.1, the single locus identified only in the rectal cancer analysis, GTEx data show that, in gastrointestinal tissues, lead SNP rs28611105 colocalises with a *cis*-eQTL coregulating expression of genes *PYGL*, *ABHD12B* and *NIN*. We reported an association between genetically predicted glycogen phosphorylase L (*PYGL*) expression and CRC risk in a transcriptome-wide association study.[47] This glycogen metabolism gene plays an important role in sustaining proliferation and preventing premature senescence in hypoxic cancer cells.[48]

At 1p31.1, identified in the colon cancer analysis, *PTGER3* encodes prostaglandin E receptor 3, a receptor for prostaglandin E2 (PGE2), a potent pro-inflammatory metabolite biosynthesised by cyclooxygenase-2 (COX-2). COX-2 plays a critical role in mediating inflammatory responses that lead to epithelial malignancies. The anti-inflammatory activity of non-steroidal anti-inflammatory drugs (NSAIDs) such as aspirin and ibuprofen operates mainly through COX-2 inhibition, and long-term NSAID use decreases CRC incidence and mortality.[49] PGE2 is required for the activation of β-catenin by Wnt in stem cells,[50] and promotes colon cancer cell growth.[51] *PTGER3* plays an important role in suppression of cell growth and its downregulation was shown to enhance colon carcinogenesis.[52]

Previous CRC GWASs had already reported allelic effect heterogeneity between tumour sites, including for 10p14, 11q23 and 18q21 but only contrasted colon and rectal tumours, without distinguishing between proximal and distal colon.[53 54] Sample size and timing of the present study enabled systematic characterisation of allelic effect heterogeneity between more refined tumour anatomical sublocations, and for a much expanded catalogue of risk variants. Our analysis revealed substantial, previously unappreciated allelic effect heterogeneity between proximal and distal CRC. Results further show that distal colon and rectal cancer have very similar germline genetic aetiologies. Our findings at several loci are consistent with CRC tumour molecular studies. Consensus molecular subtypes (CMSs), which are based on tumour gene expression, are differentially distributed between proximal and distal CRCs. The canonical CMS (CMS2) is enriched in distal CRC (56% vs 26% for proximal CRC) and is characterised by upregulation of Wnt downstream targets.[55] We found that variant associations near Wnt/β-catenin pathway genes *APC* and *CTNNB1* were confined to distal CRC. We also found that associations for variants near genes *BOC* and *FOXL1*, members of the Hedgehog signalling pathway, were confined to distal CRC, suggesting that Wnt and Hedgehog signalling may contribute more to the development of distal CRC tumours. However, pathway enrichment analyses did not provide clear evidence for differential involvement of pathways, suggesting perhaps that associations for proximal and distal CRC mostly converge on the same pathways. Pathway analysis results should, however, be interpreted taking into consideration the limitations of available approaches. Genetic variants were mapped to the nearest gene which is often not the target gene.

The precise intrinsic or extrinsic effect modifiers explaining observed allelic effect heterogeneity between anatomical subsites remain unknown and further research is needed. Short-chain fatty acids, in particular butyrate, produced by microbiota through fermentation of dietary fibre in the colon may be involved. Concentrations of butyrate, which plays a multifaceted antitumorigenic role in maintaining gut homoeostasis, are much higher in proximal colon.[56] Moreover, the known chemopreventive role of butyrate may involve modulation of signalling pathways including TGFβ and Wnt.[57] This may contribute to possible differences between anatomical segments in colorectal crypt cellular dynamics.

One limitation of our study is that we have not performed GWAS analyses of case subgroups based on more detailed anatomical sublocations. However, given current sample size, such analyses would result in reduced statistical power owing to reduced sample sizes and the aggravated multiple testing burden. As another limitation, our study was based on European-ancestry subjects and it remains to be determined whether findings are generalisable to other ancestries.

In conclusion, germline genetic data support the idea that proximal and distal colorectal cancer have partly distinct aetiologies. Our results further demonstrate that distal colon and rectal cancer have very similar germline genetic aetiologies and argue against lumping proximal and distal colon cancer in studies of aetiological factors. Future genetic studies should take into consideration differences between primary tumour anatomical subsites. A better understanding of differing carcinogenic mechanisms and neoplastic transformation risk in proximal and distal colorectum can inform the development of novel precision treatment and prevention strategies through the discovery of novel drug targets and repurposable drug candidates for treatment and chemoprevention, and improved individualised screening recommendations based on risk prediction models incorporating tumour anatomical subsite.

Author affiliations
[1]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA
[2]Division of Human Genetics, Department of Internal Medicine, The Ohio State University Comprehensive Cancer Center, Columbus, Ohio, USA
[3]Department of Preventive Medicine, Keck School of Medicine of the University of Southern California, Los Angeles, California, USA
[4]Department of Medicine, Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA
[5]Genomic Medicine Institute, Cleveland Clinic, Cleveland, Ohio, USA
[6]Department of Preventive Medicine and USC Norris Comprehensive Cancer Center, Keck School of Medicine of the University of Southern California, Los Angeles, California, USA
[7]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA
[8]Department of Medicine, University of North Carolina School of Medicine, Chapel Hill, North Carolina, USA
[9]Department of Epidemiology and Biostatistics, Imperial College London, London, UK
[10]Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
[11]UPMC Hillman Cancer Center, Pittsburgh, Pennsylvania, USA
[12]Translational Genomics Research Institute - An Affiliate of City of Hope, Phoenix, Arizona, USA
[13]Department of Radiation Sciences, Oncology Unit, Umeå University, Umeå, Sweden
[14]Public Health Division of Gipuzkoa, Health Department of Basque Country, San Sebastian, Spain
[15]Department of Biostatistics, University of Washington, Seattle, Washington, USA
[16]Laboratoire de Mathématiques Appliquées MAP5 (UMR CNRS 8145), Université Paris Descartes, Paris, France
[17]Centre for Research in Epidemiology and Population Health (CESP), Institut pour la Santé et la Recherche Médicale (INSERM) U1018, Université Paris-Saclay, Villejuif, France
[18]Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK
[19]Division of Research, Kaiser Permanente Northern California, Oakland, California, USA
[20]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, USA
[21]Department of Internal Medicine, University of Utah Health, Salt Lake City, Utah, USA
[22]Departments of Cancer Biology and Genetics and Internal Medicine, The Ohio State University, Columbus, Ohio, USA
[23]Division of Human Nutrition and Health, Wageningen University & Research, Wageningen, The Netherlands
[24]Unit of Nutrition and Cancer, Cancer Epidemiology Research Program, Catalan Institute of Oncology - IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain

[25]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA
[26]Cancer Prevention and Control Program, Catalan Institute of Oncology - IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain
[27]CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain
[28]Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain
[29]Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, Minnesota, USA
[30]Gastroenterology Department, Hospital Clínic, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), University of Barcelona, Barcelona, Spain
[31]Division of Clinical Epidemiology and Aging Research, German Cancer Research Centre (DKFZ), Heidelberg, Germany
[32]Department of Genetics, Stanford University, Stanford, California, USA
[33]Service de Génétique Médicale, Centre Hospitalier Universitaire (CHU) de Nantes, Nantes, France
[34]Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK
[35]Huntsman Cancer Institute and Department of Population Health Sciences, University of Utah Health, Salt Lake City, Utah, USA
[36]Department of Epidemiology, German Institute of Human Nutrition (DIfE), Potsdam-Rehbrücke, Germany
[37]Institut Gustave Roussy, Université Paris-Saclay, Villejuif, France
[38]Division of Preventive Oncology, German Cancer Research Centre (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany
[39]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany
[40]Institute of Cancer Research, Department of Medicine I, Medical University of Vienna, Vienna, Austria
[41]Department of Medicine I, University Hospital Dresden, Technische Universität Dresden (TU Dresden), Dresden, Germany
[42]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia
[43]Colorectal Oncogenomics Group, Genetic Epidemiology Laboratory, Department of Clinical Pathology, The University of Melbourne, Melbourne, Victoria, Australia
[44]Genomic Medicine and Family Cancer Clinic, Royal Melbourne Hospital, Melbourne, Victoria, Australia
[45]Institute for Health Research, Kaiser Permanente Colorado, Denver, Colorado, USA
[46]Division of Research, Kaiser Permanente Medical Care Program, Oakland, California, USA
[47]Behavioral and Epidemiology Research Group, American Cancer Society, Atlanta, Georgia, USA
[48]Division of Clinical Epidemiology, German Cancer Research Centre (DKFZ), Heidelberg, Germany
[49]Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA
[50]Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA
[51]Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA
[52]Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA
[53]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA
[54]Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA
[55]Division of Cancer Epidemiology, German Cancer Research Centre (DKFZ), Heidelberg, Germany
[56]Cancer Epidemiology Group, University Medical Centre Hamburg-Eppendorf, University Cancer Centre Hamburg (UCCH), Hamburg, Germany
[57]Department of Cancer Biology and Genetics and the Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio, USA
[58]Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
[59]Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, Victoria, Australia
[60]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, University of Toronto, Toronto, Ontario, Canada
[61]SWOG Statistical Center, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA
[62]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA
[63]Department of Medicine, University of Washington School of Medicine, Seattle, Washington, USA
[64]University of Hawai'i Cancer Center, Honolulu, Hawaii, USA
[65]Nutrition and Metabolism Section, International Agency for Research on Cancer, World Health Organization, Lyon, France
[66]Department of Epidemiology, School of Public Health and Institute of Health and Environment, Seoul National University, Seoul, South Korea
[67]Ontario Institute for Cancer Research, Toronto, Ontario, Canada
[68]Center for Gastrointestinal Biology and Disease, University of North Carolina, Chapel Hill, North Carolina, USA
[69]The Clalit Health Services, Personalized Genomic Service, Carmel Medical Center, Haifa, Israel
[70]Department of Community Medicine and Epidemiology, Lady Davis Carmel Medical Center, Haifa, Israel
[71]Clalit National Cancer Control Center, Haifa, Israel
[72]Department of Family Medicine, University of Virginia, Charlottesville, Virginia, USA
[73]Institute of Epidemiology, PopGen Biobank, Christian-Albrechts-University Kiel, Kiel, Germany
[74]Department of Clinical Genetics, Karolinska University Hospital, Stockholm, Sweden
[75]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden
[76]Department of Health Science Research, Mayo Clinic, Scottsdale, Arizona, USA
[77]Department of Public Health Solutions, National Institute for Health and Welfare, Helsinki, Finland
[78]Departments of Medicine and Genetics, Case Comprehensive Cancer Center, Case Western Reserve University and University Hospitals of Cleveland, Cleveland, Ohio, USA
[79]Department of Pathology, School of Medicine, Umm Al-Qura'a University, Mecca, Saudi Arabia
[80]Clinical Genetics Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, New York, USA
[81]Department of Medicine, Weill Cornell Medical College, New York, New York, USA
[82]Department of Oncologic Pathology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA
[83]Program in MPE Molecular Pathological Epidemiology, Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA
[84]Dipartimento di Medicina Clinica e Chirurgia, University of Naples Federico II, Naples, Italy
[85]Clinical Epidemiology Unit, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Newfoundland, Canada
[86]Department of Epidemiology, University of Washington, Seattle, Washington, USA
[87]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA
[88]Department of Public Health Sciences, School of Medicine, University of California Davis, Davis, California, USA
[89]Division of Epidemiology, Vanderbilt Epidemiology Center, Vanderbilt University School of Medicine, Nashville, Tennessee, USA
[90]Ruth and Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, Haifa, Israel
[91]School of Public Health, Imperial College London, London, UK
[92]Department of General Surgery, University Hospital Rostock, Rostock, Germany
[93]Department of Medicine and Epidemiology, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania, USA
[94]Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, Maryland, USA
[95]Department of Nutrition, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA
[96]Division of Laboratory Genetics, Department of Laboratory Medicine and Pathology, MayoClinic, Rochester, Minnesota, USA
[97]Hellenic Health Foundation, Athens, Greece
[98]WHO Collaborating Center for Nutrition and Health, Unit of Nutritional Epidemiology and Nutrition in Public Health, Department of Hygiene, Epidemiology and Medical Statistics, School of Medicine, National and Kapodistrian University of Athens, Athens, Greece
[99]Department of Molecular Biology of Cancer, Institute of Experimental Medicine of the Czech Academy of Sciences, Prague, Czech Republic
[100]Institute of Biology and Medical Genetics, First Faculty of Medicine, Charles University, Prague, Czech Republic
[101]Faculty of Medicine and Biomedical Center in Pilsen, Charles University, Pilsen, Czech Republic
[102]Medical Faculty, University of Heidelberg, Heidelberg, Germany
[103]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
[104]Discipline of Genetics, Memorial University of Newfoundland, St. John's, Newfoundland, Canada
[105]Department of Genome Sciences, University of Washington, Seattle, Washington, USA
[106]Department of Genetics and Genome Sciences, Case Western Reserve University School of Medicine, Case Comprehensive Cancer Center, Cleveland, Ohio, USA
[107]Department of Computer Science, Stanford University, Stanford, California, USA
[108]Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, USA

[109]Department of Preventive Medicine, USC Norris Comprehensive Cancer Center, University of Southern California Keck School of Medicine, Los Angeles, California, USA
[110]City of Hope National Medical Center, Duarte, California, USA
[111]Division of Epidemiology, Department of Population Health, New York University School of Medicine, New York, New York, USA

Twitter Daniel D Buchanan @dan_buchanan, Sergi Castellví-Bel @scastellvibel and Mazda Jenab @mazda_j

Contributors JRH, TAH, SAB, HH, JCF, SLS, DVC, JAB, AJC, BD, DD, SH, LI, VP, AP-C, LCS, FRS, MLS, AET, FJBvD, BVG, AA, DA, MHA, KA, CA-C, VA, SIB, SB, DTB, JB, HBoeing, M-CB-R, HBrenner, SBrezina, SBuch, DDB, AB-H, BJC, PTC, PC, AC, SC-B, ATC, JC-C, SJC, AdlC, DFE, DRE, EJMF, MG, SJG, WJG, GGG, PJG, WMG, JSG, AG, MJG, RWH, JH, MH, JLH, W-YH, TJH, MJ, MAJ, ADJ, TOK, CK, TK, SK, LLM, FL, CIL, LL, WL, AL, NML, SM, SDM, RLM, LM, NM, RN, KO, SO, SP, PSP, RP, PDPP, AIP, EAP, JDP, RLP, LQ, LR, GR, HSR, ER, CS, RES, DS, MS, CMT, SNT, DCT, AT, CMU, KV, PV, LV, VV, KW, SJW, EW, AW, MOW, AHW, GRA, DAN, PCS, AK, GC, SBG, LH, VM, RBH, PAN and UP conceived and designed the study. JRH, TAH, SAB, SLS, DVC, SC, CQ, YL, RB, HMK, DML, FRS, BB, KRC, W-LH, Y-RS, AK, LH and UP analysed the data. JRH, TAH, HH, JCF, JAB, AJC, BD, SH, LI, HMK, VP, AP-C, LCS, MLS, AET, FJBvD, BVG, AA, DA, MHA, KA, CA-C, VA, MCB, SIB, SB, DTB, JB, HBoeing, M-CB-R, HBrenner, SBrezina, SBuch, DDB, AB-H, BJC, PTC, PC, AC, SC-B, ATC, JC-C, SJC, AdlC, DFE, DRE, EJMF, MG, SJG, WJG, GGG, PJG, WMG, JSG, AG, MJG, RWH, JH, MH, JLH, W-LH, W-YH, TJH, MJ, MAJ, ADJ, TOK, CK, TK, SK, LLM, FL, CIL, LL, WL, AL, NML, SM, SDM, RLM, LM, NM, RN, KO, SO, SP, PSP, RP, PDPP, AIP, EAP, JDP, RLP, LQ, LR, GR, HSR, ER, CS, RES, MS, Y-RS, CMT, SNT, DCT, AT, CMU, KV, PV, LV, VM, KW, SJW, EW, AW, MOW, AHW, GRA, DAN, PCS, AK, GC, SBG, VM, RBH, PAN and UP contributed reagents/materials/analysis tools. JRH, TH and UP wrote the first draft. All authors reviewed the manuscript for intellectual content and approved the final version of the manuscript. UP supervised the study.

Funding This work was supported by grants from the National Cancer Institute (NCI), National Institutes of Health (NIH), US Department of Health and Human Services (U01 CA164930, U01 CA137088, R01 CA059045, R21 CA191312, R01 CA201407, P30 CA015704). Genotyping services were provided by the Center for Inherited Disease Research (CIDR; X01-HG008596 and X01-HG007585). CIDR is fully funded through a federal contract from the NIH to the Johns Hopkins University, contract HHSN268201200008I. The full list of funding and acknowledgements can be found in the supplemental file.

Disclaimer Where authors are identified as personnel of the International Agency for Research on Cancer/WHO, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/WHO.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public controlled access repository. All genotype data analyzed in this study have been previously published and have been deposited in the database of Genotypes and Phenotypes (dbGaP), which is hosted by the National Center for Biotechnology Information (NCBI) of the US National Institutes of Health (NIH), under accession numbers phs001415.v1.p1, phs001315.v1.p1, phs001078.v1.p1, and phs001903.v1.p1. The UK Biobank resource was accessed through application number 8614. Bioinformatic analyses included public, open access colorectal epigenomic data that were retrieved from the NCBI Gene Expression Omnibus (GEO) database under accession numbers GSE77737 and GSE36401. For all above datasets embargo release dates have passed.

ORCID iDs
Jeroen R Huyghe http://orcid.org/0000-0001-6027-9806
Hermann Brenner http://orcid.org/0000-0002-6129-1572
Daniel D Buchanan http://orcid.org/0000-0003-2225-6675
Andrew T Chan http://orcid.org/0000-0001-7284-6767
Andrea Gsur http://orcid.org/0000-0002-9795-1528
Jochen Hampe http://orcid.org/0000-0002-2421-6127
Michael Hoffmeister http://orcid.org/0000-0002-8307-3197
Wen-Yi Huang http://orcid.org/0000-0002-4440-3368
Mazda Jenab http://orcid.org/0000-0002-0573-1852
Neil Murphy http://orcid.org/0000-0003-3347-8249
Ulrike Peters http://orcid.org/0000-0001-5666-9318

REFERENCES

1 American Cancer Society. Cancer statistics center. Available: http://cancerstatistics center.cancer.org [Accessed 21 Apr 2020].
2 Bufill JA. Colorectal cancer: evidence for distinct genetic categories based on proximal or distal tumor location. *Ann Intern Med* 1990;113:779–88.
3 Iacopetta B. Are there two sides to colorectal cancer? *Int J Cancer* 2002;101:403–8.
4 Carethers JM. One colon lumen but two organs. *Gastroenterology* 2011;141:411–2.
5 Yamauchi M, Lochhead P, Morikawa T, *et al*. Colorectal cancer: a tale of two sides or a continuum? *Gut* 2012;61:794–7.
6 Lichtenstein P, Holm NV, Verkasalo PK, *et al*. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* 2000;343:78–85.
7 Schmit SL, Edlund CK, Schumacher FR, *et al*. Novel common genetic susceptibility loci for colorectal cancer. *J Natl Cancer Inst* 2019;111:146–57.
8 Huyghe JR, Bien SA, Harrison TA, *et al*. Discovery of common and rare genetic risk variants for colorectal cancer. *Nat Genet* 2019;51:76–87.
9 Law PJ, Timofeeva M, Fernandez-Rozadilla C, *et al*. Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nat Commun* 2019;10:2154.
10 Lu Y, Kweon S-S, Tanikawa C, *et al*. Large-Scale genome-wide association study of East Asians identifies loci associated with risk for colorectal cancer. *Gastroenterology* 2019;156:1455–66.
11 Peters U, Jiao S, Schumacher FR, *et al*. Identification of genetic susceptibility loci for colorectal tumors in a genome-wide meta-analysis. *Gastroenterology* 2013;144:799–807.
12 Schumacher FR, Schmit SL, Jiao S, *et al*. Genome-Wide association study of colorectal cancer identifies six new susceptibility loci. *Nat Commun* 2015;6:7138.
13 McCarthy S, Das S, Kretzschmar W, *et al*. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;48:1279–83.
14 Delaneau O, Howie B, Cox AJ, *et al*. Haplotype estimation using sequencing reads. *Am J Hum Genet* 2013;93:687–96.
15 Das S, Forer L, Schönherr S, *et al*. Next-Generation genotype imputation service and methods. *Nat Genet* 2016;48:1284–7.
16 Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010;26:2190–1.
17 Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* 2013;381:1371–9.
18 Burnham KP, Anderson DR. Multimodel inference: understanding AIC and BIC in model selection. *Sociol Methods Res* 2004;33:261–304.
19 Yee TW. The VGAM Package for Categorical Data Analysis. *J Stat Softw* 2010;32.
20 Lamparter D, Marbach D, Rueedi R, *et al*. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Comput Biol* 2016;12:e1004714.
21 Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28:27–30.
22 Croft D, O'Kelly G, Wu G, *et al*. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* 2011;39:D691–7.
23 Nishimura D. BioCarta. *Biotech Software & Internet Report* 2001;2:117–20.
24 Traylor M, Markus H, Lewis CM. Homogeneous case subgroups increase power in genetic association studies. *Eur J Hum Genet* 2015;23:863–9.
25 Raptis S, Mrkonjic M, Green RC, *et al*. MLH1 -93G>A promoter polymorphism and the risk of microsatellite-unstable colorectal cancer. *J Natl Cancer Inst* 2007;99:463–74.
26 Mrkonjic M, Roslin NM, Greenwood CM, *et al*. Specific variants in the MLH1 gene region may drive DNA methylation, loss of protein expression, and MSI-H colorectal cancer. *PLoS One* 2010;5:e13314.
27 Cunningham JM, Christensen ER, Tester DJ, *et al*. Hypermethylation of the hMLH1 promoter in colon cancer with microsatellite instability. *Cancer Res* 1998;58:3455–60.
28 Avram D, Califano D. The multifaceted roles of Bcl11b in thymic and peripheral T cells: impact on immune diseases. *J Immunol* 2014;193:2059–65.
29 Punwani D, Zhang Y, Yu J, *et al*. Multisystem anomalies in severe combined immunodeficiency with mutant Bcl11b. *N Engl J Med* 2016;375:2165–76.
30 Kadoch C, Hargreaves DC, Hodges C, *et al*. Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. *Nat Genet* 2013;45:592–601.

31 Gutierrez A, Kentsis A, Sanda T, *et al*. The BCL11B tumor suppressor is mutated across the major molecular subtypes of T-cell acute lymphoblastic leukemia. *Blood* 2011;118:4169–73.
32 Neumann M, Vosberg S, Schlee C, *et al*. Mutational spectrum of adult T-ALL. *Oncotarget* 2015;6:2754–66.
33 Sakamaki A, Katsuragi Y, Otsuka K, *et al*. Bcl11b SWI/SNF-complex subunit modulates intestinal adenoma and regeneration after γ-irradiation through Wnt/β-catenin pathway. *Carcinogenesis* 2015;36:622–31.
34 Kajiho H, Saito K, Tsujita K, *et al*. RIN3: a novel Rab5 GEF interacting with amphiphysin II involved in the early endocytic pathway. *J Cell Sci* 2003;116:4159–68.
35 Kajiho H, Sakurai K, Minoda T, *et al*. Characterization of RIN3 as a guanine nucleotide exchange factor for the Rab5 subfamily GTPase Rab31. *J Biol Chem* 2011;286:24364–73.
36 Jones MF, Hara T, Francis P, *et al*. The CDX1-microRNA-215 axis regulates colorectal cancer stem cell differentiation. *Proc Natl Acad Sci U S A* 2015;112:E1550–8.
37 Guo R-J, Huang E, Ezaki T, *et al*. Cdx1 inhibits human colon cancer cell proliferation by reducing beta-catenin/T-cell factor transcriptional activity. *J Biol Chem* 2004;279:36865–75.
38 Rasinperä H, Forsblom C, Enattah NS, *et al*. The C/C-13910 genotype of adult-type hypolactasia is associated with an increased risk of colorectal cancer in the Finnish population. *Gut* 2005;54:643–7.
39 World Cancer Research Fund/American Institute for Cancer Research. Continuous update project expert report 2018. diet, nutrition, physical activity and colorectal cancer. Available: dietandcancerreport.org
40 Campbell CD, Ogburn EL, Lunetta KL, *et al*. Demonstrating stratification in a European American population. *Nat Genet* 2005;37:868–72.
41 Truty MJ, Lomberk G, Fernandez-Zapico ME, *et al*. Silencing of the transforming growth factor-beta (TGFbeta) receptor II by Kruppel-like factor 14 underscores the importance of a negative feedback mechanism in TGFbeta signaling. *J Biol Chem* 2009;284:6291–300.
42 Small KS, Hedman AK, Grundberg E, *et al*. Identification of an imprinted master trans regulator at the Klf14 locus related to multiple metabolic phenotypes. *Nat Genet* 2011;43:1040–4.
43 Small KS, Todorčević M, Civelek M, *et al*. Regulatory variants at Klf14 influence type 2 diabetes risk via a female-specific effect on adipocyte size and body composition. *Nat Genet* 2018;50:572–80.
44 Stacey SN, Sulem P, Masson G, *et al*. New common variants affecting susceptibility to basal cell carcinoma. *Nat Genet* 2009;41:909–14.
45 Eikesdal HP, Becker LM, Teng Y, *et al*. BMP7 Signaling in *TGFBR2*-Deficient Stromal Cells Provokes Epithelial Carcinogenesis. *Mol Cancer Res* 2018;16:1568–78.
46 Motoyama K, Tanaka F, Kosaka Y, *et al*. Clinical significance of BMP7 in human colorectal cancer. *Ann Surg Oncol* 2008;15:1530–7.
47 Bien SA, Su Y-R, Conti DV, *et al*. Genetic variant predictors of gene expression provide new insight into risk of colorectal cancer. *Hum Genet* 2019;138:307–26.
48 Favaro E, Bensaad K, Chong MG, *et al*. Glucose utilization via glycogen phosphorylase sustains proliferation and prevents premature senescence in cancer cells. *Cell Metab* 2012;16:751–64.
49 Jänne PA, Mayer RJ. Chemoprevention of colorectal cancer. *N Engl J Med* 2000;342:1960–8.
50 Goessling W, North TE, Loewer S, *et al*. Genetic interaction of PGE2 and Wnt signaling regulates developmental specification of stem cells and regeneration. *Cell* 2009;136:1136–47.
51 Castellone MD, Teramoto H, Williams BO, *et al*. Prostaglandin E2 promotes colon cancer cell growth through a Gs-axin-beta-catenin signaling axis. *Science* 2005;310:1504–10.
52 Shoji Y, Takahashi M, Kitamura T, *et al*. Downregulation of prostaglandin E receptor subtype EP3 during colon cancer development. *Gut* 2004;53:1151–8.
53 Tenesa A, Farrington SM, Prendergast JGD, *et al*. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet* 2008;40:631–7.
54 Tomlinson IPM, Webb E, Carvajal-Carmona L, *et al*. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet* 2008;40:623–30.
55 Guinney J, Dienstmann R, Wang X, *et al*. The consensus molecular subtypes of colorectal cancer. *Nat Med* 2015;21:1350–6.
56 Tan J, McKenzie C, Potamitis M, *et al*. The role of short-chain fatty acids in health and disease. *Adv Immunol* 2014;121:91–119.
57 McNabney SM, Henagan TM. Short chain fatty acids in the colon and peripheral tissues: a focus on butyrate, colon cancer, obesity and insulin resistance. *Nutrients* 2017;9:9.